Error Detection in Panoramic Videos: a Pairwise Assessment within Stitching

Sandra Nabil, Frédéric Devernay, James L. Crowley

ERROR DETECTION IN PANORAMIC VIDEOS: A PAIRWISE ASSESSMENT WITHIN STITCHING

Sandra Nabil, Frederic Devernay, James Crowley

INRIA Grenoble Alpes

ABSTRACT

One way to provide realistic immersive VR content relies on producing high-quality panoramic videos. These videos are usually produced using multiple cameras with different optical centers and which may not be perfectly synchronized This results in spatial and temporal artifact, even though the blending algorithm strives to reduce them. In this paper, we devise a method that detects potential visual artifacts, based on existing view synthesis quality metrics. The method works by computing pair-wise quality at each blending step and fusing them to produce a global map of potential errors. To get a more accurate prediction, we develop a mask that is then applied to the error map and therefore accentuates the defects on the blending cutting line. Results show that the calculated distortion map succeeds to identify visual artifacts in panoramas which can help design better solutions to this problem in the future.

***Index Terms*—** panoramic videos, image blending, quality metrics, error prediction, parallax error

1. INTRODUCTION

Panoramic videos are becoming an important tool for providing immersion in virtual reality (VR) environments. The richness of 360 panoramic videos makes it possible to capture outstanding scenes where the user can experience presence by looking through a head-mounted display and turning around without missing anything of the surrounding events. However, this realism can easily be broken if the user spots any visual artifact. One of the most disturbing visual error phenomena falls in the category of parallax errors which appear in the form of discontinuities, deformations or ghosting. This kind of distortion is nearly unavoidable in panoramic video capture that is usually created by the use of a panoramic rig, which assembles multiple cameras each covering a large field of view with a certain overlap with one or more other cameras. Although there have been several attempts to prevent this kind of error [1, 2], it is still far from being entirely resolved; and even detecting those errors remains a challenge. Figure 1 shows examples of parallax errors extracted from a panoramic video frame captured by five cameras.

Many researchers have already presented methods to solve the problem either by compensating the parallax error [1] or by optimizing the projected images in a mesh grid fashion [2] as we will see in more details in the next section. Although many of these attempts reduced the number or amplitude of visual artifacts, every one of them has unexpected failure cases and is thus not suited for general use. Unlike these methods, we only focus on understanding those errors and detecting their locations in the panorama. We thought that understanding how much human vision is sensitive to these defects can help finding a solution to avoid as many of them as possible. Hence, we redirect our focus from problem solving to problem understanding and detection.

Fig. 1. Examples of parallax errors in panoramic videos. From left to right, the deformation of the person's head, the misalignment at the top of the building and the ghosting of the statue and the misalignment and loss of part of the chair in the last image. The first is taken from results by [1] which includes a parallax compensation step prior to the final multi-band blending. The second and third are produced using the open source software Hugin [3] with multi-band blending and no parallax compensation.

In this paper, we provide a new method to identify the location as well as the significance of visual artifacts in a panoramic video frame. We make use of an image quality metric originally designed to assess the quality of depth-image based rendering (DIBR consists by definition in synthesizing new viewpoints for an image and a depth map or a stereo pair) for a new purpose which is error prediction embedded within the image stitching process, regardless of the blending algorithm used. The metric used, known as visual synthesis quality assessment (VSQA) metric [4] is an extension of the well-known structural similarity (SSIM [5]) map, with the addition of three visibility maps that are used as weights for the SSIM map. These weighting maps are inspired from the human perception and sensitivity of visual

errors. We apply this measurement using the blending order where we compare each pair of overlapping regions to each other, with and without parallax compensation. This allows predicting errors prior to the final step of image blending. Afterwards, we apply a mask that is calculated on the blending cut between images to give more weight to the errors appearing near the transitional area between a pair of images. Results show that this approach successfully spots potential errors giving more importance to the ones that are more visible according to human visual system (HVS).

The rest of the paper is organized as follows: we first provide an overview of related work and a background necessary for understanding our approach. In section 2, we elaborate on the new suggested approach and show results in section 3. Finally, we conclude the proposed work and highlight future research directions.

2. BACKGROUND AND RELATED WORK

The following section is divided in two parts that are both essential to the comprehension of our work. In section 2.1, we give an overview of the typical image stitching process and the variations that are specific to video stitching. We discuss in section 2.2 recent work on quality metrics for image and video, and focus on the specific case of quality metrics for view synthesis, which is the core of our error prediction method.

2.1. Image and video stitching

Image stitching is a multi-step process that aims to produce a wide-angle or panoramic view of a scene from a number of overlapping images or videos taken from the same viewpoint [6, 7]. The method first extract feature points in each image that are compared afterwards with key-points in other images to determine pairwise matches. It then tries to establish a mathematical relationship between each image or video pair using model fitting methods such as RANSAC [8]. This can be used to calculate the rotation between images, which corresponds to a 2D transformation between images if all cameras have the same optical center. Once we have the relative positions of each camera, an appropriate projection surface is chosen and each image is projected individually into it. Finally, an iterative blending method is applied on the whole set of images to create one seamless panorama.

This method is well-established and is considered the baseline used in most commercial and open-source image stitching software as well as popular computer vision libraries such as OpenCV. Although this method succeeds to create visually pleasant panoramic photos that have minimal noticeable artifacts, it assumes very small to no translation between the optical centers of the cameras, so that there is no parallax. In the process of taking panoramic videos the cameras cannot physically share a common optical center, which may result in what we call parallax errors, caused by the difference in position of the camera centers between two or more viewing points to the same scene. In the images, parallax error appear in the form of spatial or temporal image discontinuities, double images (also called ghosting) or deformations.

For this reason, algorithms for panoramic image creation may not be directly applicable in the case of panoramic videos. To resolve this issue, research has focused on compensating parallax errors in different ways. Perazzi et al. [1] proposed a solution consisting in parallax estimation between overlapping areas using optical flow, followed by a calculation of error between one view, the other warped towards it and their distance from the fused image in a patch-based approach. This error is used to calculate an optimal warp order, which goal is to minimize parallax error. The results seem promising, however the error calculation is very expensive and the method fails for cases of high displacement and motion blur. It also highly depends on the choice of the reference frame.

Lee et al. [2] propose to project images from different viewpoints onto a deformable 3D sphere and optimize locally the resolution based on visual saliency. This approach simplifies the problem into a simple and fast mesh deformation problem, however it is vulnerable to errors in the calibration step. It also requires user input to specify important parts which are given high resolution but which are not calculated with respect to the final panorama resolution. Lin et al. [9] solve the problem using a 3D reconstruction and image overlaying rather than blending. This method is useful in the case of hand-held cameras. However, 3D reconstruction from video is costly and image overlaying requires a choice of one view which may cause temporal artifacts if there is camera or scene motion. What is common about all of these methods, is that they all provide a solution in an attempt to directly attenuate these errors, without an actual explanation of the reason why and at which step these errors occur.

2.2. Image and video quality metrics

In order to assess the quality of panoramic images, one might think that using ordinary image quality metrics can do the job. However, the fact that there are multiple source/reference images to a single panoramic output with no ground truth to compare makes the problem more complicated. In addition, the processing that occur on those input images includes various modifications to the original images, causing not only photo-metric distortions but more importantly geometric ones, as mentioned in the previous section. Few works deal already with the panoramic images assessment and lots of them are quite recent, such as [10] who describe a method that assess geometric image quality for panoramas based on the variation of flow field between two given scenes weighted by a salience map in addition to a structure histogram. The method seems effective, however it depends on the per-pixel

motion field which might be erroneous itself. Others have focused on the user experience such as in [11] who provided a subjective quality experiment, while authors in [12] have provided an objective measurement based on the Peak signal-to-noise ratio PSNR metric and salience maps to assess areas that catch user attention in a virtual reality environment.

More work has been done to establish quality metrics for a field closely related to panorama creation, which is novel view synthesis or image-based rendering. The nature of this type of methods resembles image stitching in many ways, since it combines multiple views into a newly generated image, which involves a process of image warping same as in image stitching. Conze et al. [4] have designed a metric for novel view synthesis called "view synthesis quality assessment" (VSQA) metric, which is based on the structural similarity SSIM metric weighted by 3 visibility maps that reflect texture, orientation and contrast features in the image. Details are provided in the next section.

View synthesis quality assessment VSQA metric is an objective image quality metric designed for the special case of novel view synthesis, based on human perception sensitivity to artifacts. According to the authors, the human vision system (HVS) is mostly sensitive to local image variations in texture, gradient orientation diversity and high contrast areas. Therefore, they extend the well-established image metric structural similarity image index (SSIM) with three visibility weighting maps, that increase or decrease the distortion value depending on its visual saliency. They choose their reference as one view of their original images and the synthesized is the other view warped towards it as explained in the figure below. Later on, Battisti et al. [13] also suggested an interesting approach based on the comparison between statistical features of wavelet transforms as well as a method for skin detection. We used to choose the former method for its simplicity and the idea that it serves our goal sufficiently.

3. PROPOSED APPROACH

In this section, we explain our approach, whose goal is to provide a quality prediction for the panorama before the actual blending takes place. We choose to do our error calculation prior to blending for three main reasons: first, although blending strives to remove some artifacts, it is a blind method that can introduce new artifacts by removing parts of objects or mistakenly erasing something that is not actually an error. Second, once images have been blended into the final panorama, it is very difficult to recover the original images, which are as the name of the method implies, blended and mixed together in the overlapping areas, therefore post-processing to correct defects will also be difficult. Finally, to detect misalignment and discontinuities, it is essential to compare the structural dissimilarities between intersecting views, which is only available prior to blending.

Given a number of input views, we go through the stitching steps explained in section 2.1 without proceeding to the final step of blending. We examine the differences between pairs of views in two cases that are demonstrated in figure 2:

1. Non-warped views in the overlapping regions in the order in which they appear in blending.
2. One unchanged view and the other warped towards it in the overlapping regions in an optimal order calculated as suggested in [1].

As explained in the previous sections, we use the VSQA quality metric [4], which was designed for DIBR/novel view synthesis, with a new goal, which is error prediction and identification in panoramas. The VSQA metric is defined as follows:

$$VSQA(i,j) = \mathrm{dist}(i,j).[W_t]^{\delta}.[W_o]^{\epsilon}.[W_c]^{\zeta}. \quad (1)$$

where $dist$ is the chosen metric, in this case SSIM [5], calculated between a reference view and a synthesized view. This metric is weighted by 3 maps, each representing a type of local feature to which the human eye is most sensitive. Below is a list of these terms (please refer to the original paper [4] for more details):

The texture-based visibility weighting map W_t which compares the gradient of a pixel with respect to its neighbors.

$$V_t(i,j) = \frac{1}{N^2} \sum_{l=i-\lfloor \frac{N}{2} \rfloor}^{i+\lfloor \frac{N}{2} \rfloor} \sum_{k=j-\lfloor \frac{N}{2} \rfloor}^{j+\lfloor \frac{N}{2} \rfloor} w_{l,k}\mathrm{grad}[l,k], \quad (2)$$

$$W_t(i,j) = \frac{2V_t(i,j) - \min V_t}{\max V_t - \min V_t}. \quad (3)$$

The orientation-based visibility weighting map W_o which calculates the diversity of the gradient orientation of a pixel with respect to its neighbors.

$$V_o(i,j) = min_q[\frac{1}{N^2} \sum_{l=i-\lfloor \frac{N}{2} \rfloor}^{i+\lfloor \frac{N}{2} \rfloor} \sum_{k=j-\lfloor \frac{N}{2} \rfloor}^{j+\lfloor \frac{N}{2} \rfloor} w_{l,k}\min[(\theta(l,k)-\theta_q)^2, (\theta(l,k)+\pi-\theta_q)^2]], \quad (4)$$

$$W_o(i,j) = \frac{2V_o(i,j) - \min V_o}{\max V_o - \min V_o}. \quad (5)$$

The contrast-based visibility weighting map W_c which evaluates the contrast of a pixel with respect to its neighbors.

$$V_c(i,j) = \frac{1}{N^2} \sum_{l=i-\lfloor \frac{N}{2} \rfloor}^{i+\lfloor \frac{N}{2} \rfloor} \sum_{k=j-\lfloor \frac{N}{2} \rfloor}^{j+\lfloor \frac{N}{2} \rfloor} w_{l,k}|\mathrm{Lum}(l,k) - Lum(i,j)|, \quad (6)$$

$$W_c(i,j) = \frac{2V_c(i,j) - \max V_c}{\min V_c - \max V_c}. \quad (7)$$

In all of the 3 equations, N is the window size and $w_{l,k}$ is a Gaussian weight.

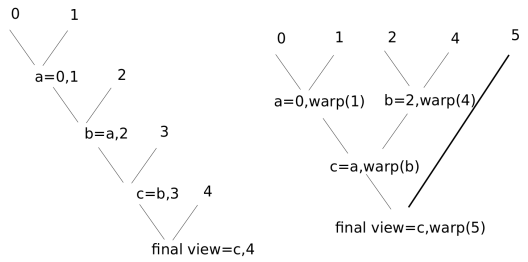

Fig. 2. Different blending orders used in the current work. Left is progressive blending in the order of image appearance used in Hugin [3]. Right is optimal warp order proposed by the authors of [1] who apply a parallax compensation.

3.1. Creation of a composite VSQA map

The VSQA map explained above is a similarity metric between two images, where one is the reference and the other synthesized or processed. In our case, we do not have a single original image and a processed output, however we have N input views and one final output, so we build our error map, by comparing each pair of images in the same order of the blending tree shown in 2 and creating one final composite map. Consider N views at a time t, after calculating pair-wise matches $P_n(i,j)$, for each pair I_i and I_j, we calculate the region of overlap $I_i \cap I_j$ and we compute VSQA metric between the region of interest in each view δI_i and δI_j.
We finally calculate the equation 8 to generate a global map for the whole panorama:

$$VSQA_{global(}i,j) = max_{i,j} VSQA_{i,j}(\delta I_i, \delta I_j). \quad (8)$$

Where i, j represent pixel location.
We test another case where we choose one view to be warped towards the other and in that case the unchanged view is considered the reference. This will change 8 to:

$$VSQA_{global}(i,j) = max_{i,j} VSQA_{i,j}(\delta I_i, warp(\delta I_j)). \quad (9)$$

Afterwards, we normalize the output map globally and we obtain what is shown in the results section in the second rows of figure 4 for a panorama without parallax compensation and 5 for the parallax compensation case.

3.2. Applying a weighting mask to VSQA map

The steps described above permit to give a global prediction of all possible areas where parallax errors can occur by comparing pairs of overlapping regions and identifying structural differences weighted by masks that enforce distortions in areas that are more salient with respect to human perception. However, as mentioned earlier, the blending step aims mainly to remove as many of these errors as possible, though it does not succeed in all the cases. The multi-band blend described in [14] usually uses a Voronoi mask that chooses the blending line to be irregular and therefore more difficult to notice a line between boundaries. But still there will be more probability to see errors around this line where one can imagine it as a pathway between both images, so we assume that the closer the pixels are to that boundary line, the more visible it is. Based on this assumption, we propose to create a weighting mask around this blending edge, which will give more weight to the pixels that fall onto this line and decreases gradually the more we go farther away. Within the same iterations over pair-wise matches as described in the previous sub-section, for a pair of views I_i and I_j, we calculate the Voronoi seam cut which produces a mask for each view M_i and M_j that determine the cutting line between both views. We are also interested only in the region of intersection between the two images, so we use the sub-masks δM_i and δM_j. In order to create the desired mask which gives weight to the errors on the blending cut, we calculate a distance transform from that line for each of the latter sub-masks, we then calculate a common mask that will be applied to the resulting VSQA as the OR between δM_i and δM_j and we get a mask M_{blend} that we normalize between 0 and 1 as shown in 3. We multiply this mask to our VSQA computed at each step in order to enforce errors at the region where the transition between images takes place and attenuate errors farther away from this boundary as described in equation 10. We call this measure MVSQA.

Fig. 3. Example of the suggested mask created around the boundary of the blending line

$$MVSQA = M_{blend}.VSQA. \quad (10)$$

We generate the global MVSQA with the same process used to calculate the composite VSQA as described previously.

4. EXPERIMENTAL RESULTS

In order to test our method, we used datasets provided by Perazzi [1] for their work on panoramic videos. We picked the *Opera* dataset, created from 5 input views, which has visual artifacts due to parallax as well as non-synchronized cameras. We also took our own panoramas using a 3-camera rig formed of Panasonic GH2 cameras with 20mm lens. Video

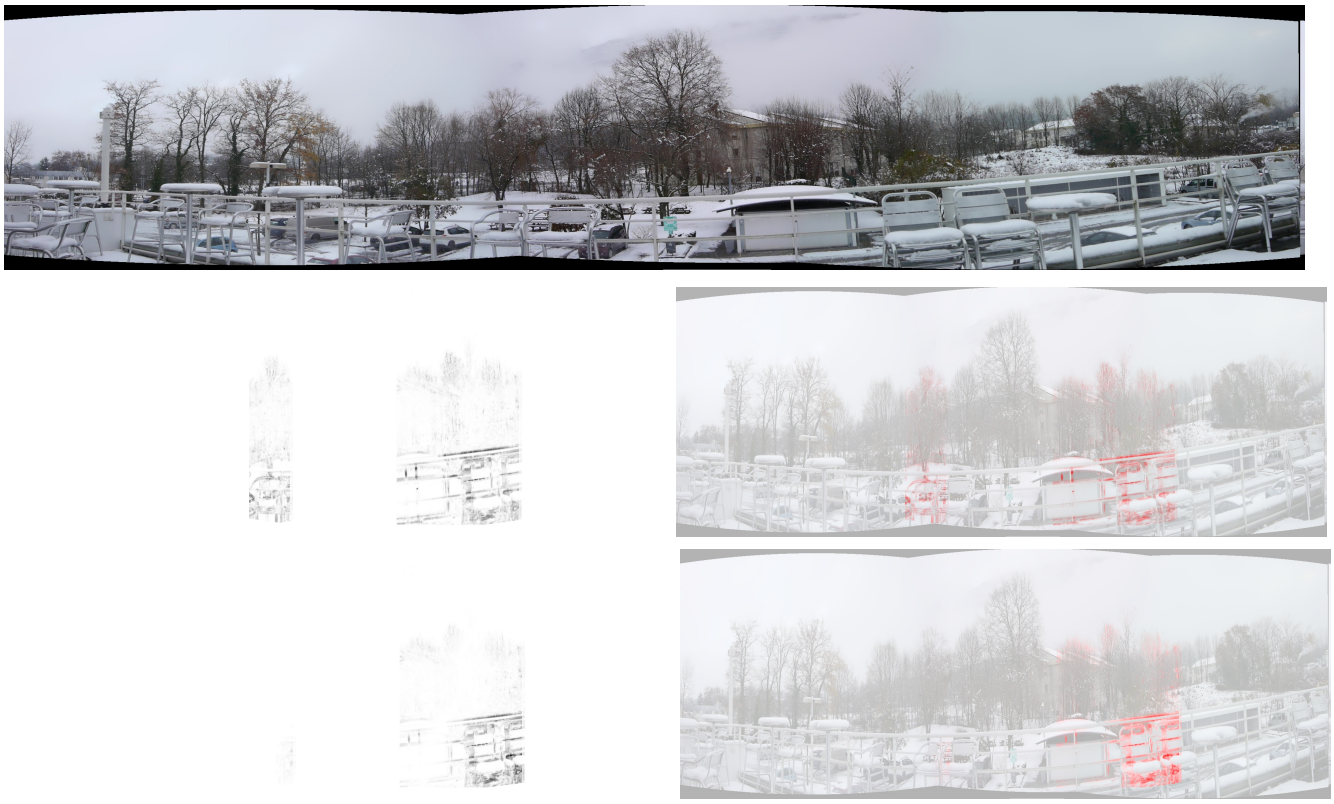

Fig. 4. The top row shows a panoramic scene of snow created by Hugin [3] and its corresponding VSQA (equation 8) and masked VSQA

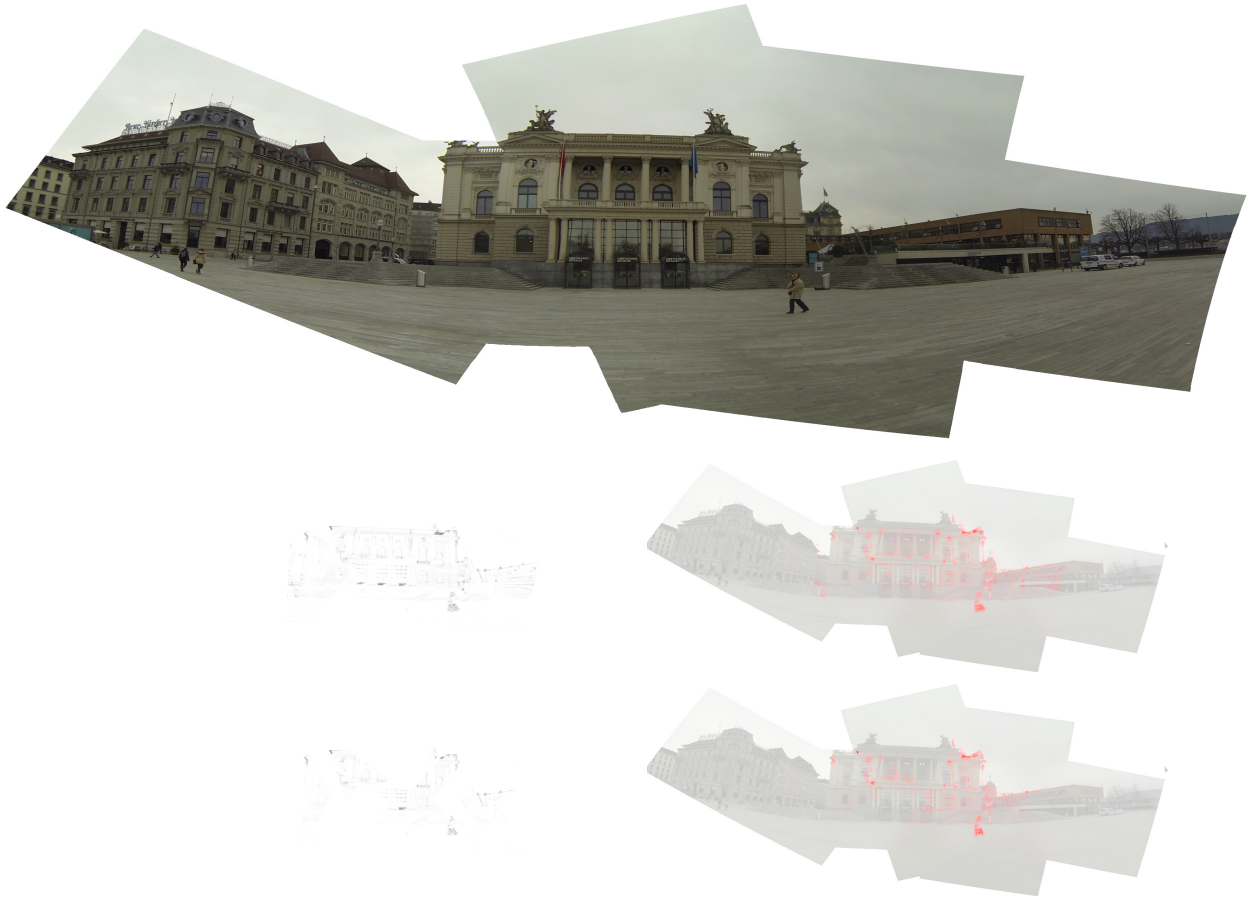

Fig. 5. The top row shows an opera panorama created using [1] and its corresponding VSQA (equation 9) and masked VSQA

frames were generated both using the open source software Hugin [3] for panorama creation, with multi-band blending and two types of mask generation graph cut and nearest feature. The authors [1] have also provided the output videos of their algorithm which has a parallax compensation using optical flow and optimal warp order. Results for the Opera dataset are shown in figure 5 and for the snow scenery in figure 4.

The results show a promising prediction for zones of potential errors not only spatially but across the whole sequence. Repeating the process for some key-frames in the video, can show which errors persist and which appear sporadically. It can also be noticed that the error seems concentrated in the right middle part of the panorama which contains four out of the five views overlapping, which is a more complicated part to stitch and one that can accumulate errors. The figures on the right show the degree of erroneous of parallax compensated frames, these show a lower values of pixels since it is a step that reduces differences between views, however, still in the region of concentrated overlaps, there is still higher values of error. The other error maps are also an indicator, however they seem to exaggerate the errors in the background where VSQA succeeds to attenuate given the human perception facts that they are less likely to have artifacts given the small contrast. Parallax error maps are also shown in the case of parallax compensation, to compare the metric used for optimal warping generation in [1].

5. CONCLUSION

We presented a method for panoramic video quality prediction integrated within the stitching process. Our experiments show it can be beneficial to compare images before blending them all together, as it can show all potential artifact location, including the ones that may be removed by the blending, which usually appear with lower intensities in the error map. The application of our calculated mask filters the errors further, which shows errors that persist after blending. We continue to work on this approach in the goal of adding a temporal factor that will help to assess a video globally rather than frame by frame. We also believe that this pre-evaluation can guide us to make a content-aware blending that is able to avoid visual artifacts that other methods have failed to resolve in complex situations.

6. REFERENCES

[1] Federico Perazzi, Alexander Sorkine-Hornung, Henning Zimmer, Peter Kaufmann, Oliver Wang, S Watson, and Markus H Gross, "Panoramic video from unstructured camera arrays," *Comput Graph Forum*, vol. 34, no. 2, 2015.

[2] Jungjin Lee, Bumki Kim, Kyehyun Kim, Younghui Kim, and Junyong Noh, "Rich360: Optimized spherical representation from structured panoramic camera arrays," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 63:1–63:11, July 2016.

[3] Hugin: Panorama photo stitcher, "Open source software," http://hugin.sourceforge.net/.

[4] Pierre-Henri Conze, Philippe Robert, and Luce Morin, "Objective View Synthesis Quality Assessment," in *Stereoscopic Displays and Applications*, SPIE, Ed., San Francisco, United States, Jan. 2012, vol. 8288 of *Proc SPIE*, pp. 8288–56.

[5] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[6] Rick Szeliski, "Image alignment and stitching: A tutorial," Tech. Rep., October 2004.

[7] Matthew Brown and David G. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, Aug 2007.

[8] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[9] Kaimo Lin, Shuaicheng Liu, Loong-Fah Cheong, and Bing Zeng, "Seamless video stitching from hand-held camera inputs," *Computer Graphics Forum*, vol. 35, no. 2, pp. 479–487, 2016.

[10] Luyu Yang, Zhigang Tan, Zhe Huang, and Gene Cheung, "A content-aware metric for stitched panoramic image quality assessment," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] Bo Zhang, Junzhe Zhao, Shu Yang, Yang Zhang, Jing Wang, and Zesong Fei, "Subjective and objective quality assessment of panoramic videos in virtual reality environments," *2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, vol. 00, pp. 163–168, 2017.

[12] M. Xu, C. Li, Z. Wang, and Z. Chen, "Visual Quality Assessment of Panoramic Video," *ArXiv e-prints*, Sept. 2017.

[13] Federica Battisti, Emilie Bosc, Marco Carli, Patrick Le Callet, and Simone Perugia, "Objective image quality assessment of 3D synthesized views," *Image Commun.*, vol. 30, no. C, pp. 78–88, Jan. 2015.

[14] Peter J. Burt and Edward H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Trans. Graph.*, vol. 2, no. 4, pp. 217–236, Oct. 1983.